

Les trans-applications

Ce jeu de transparents à pour objectif de vous montrer des applications développées au RALI et dont le but est de faciliter la vie du traducteur sans l'évincer totalement du processus de traduction...

Le projet TransType : traduction probabiliste embarquée dans un éditeur de texte

Le projet TransSearch : mémoire de traduction interactive: la saga.

EBMT : une étude sur les systèmes d'interrogation de mémoire de traduction

TransType: un aide à la rédaction de traduction

Avant TransType

traduction automatique + post-édition:

Src	The problem of famine in developing countries
Tgt	Le problème de la famine dans en voie de développement les pays

↪ **Limites:** *les traducteurs détestent corriger l'ordinateur*

traduction interactive:

Il mange sa soupe de bon appétit

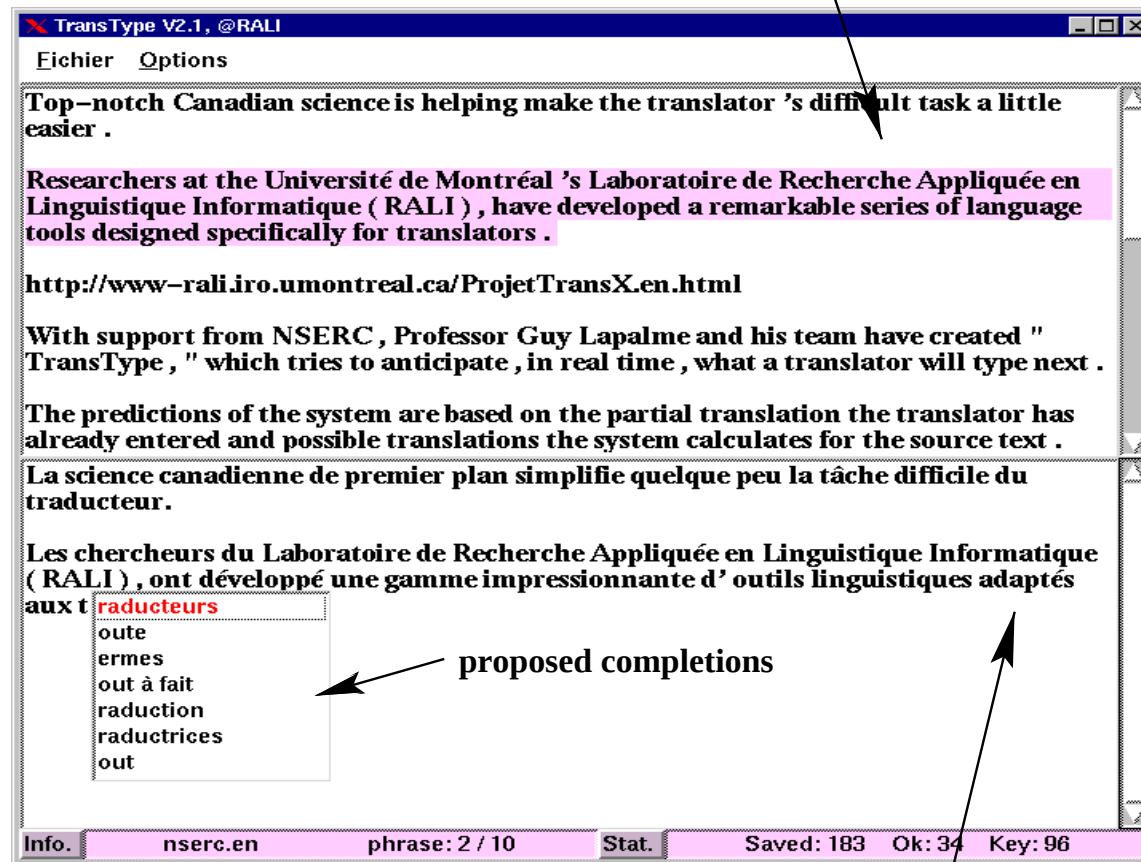
⇒ Please select one of these interpretations:

- 1) GS(PRON).GV(VB.GN(ART.NC.PREP.ADJ.NC))
- 2) GS(PRON).GV(VB.GN(ART.NC)).GC(PREP.ADJ.NC)

↪ **Limite:** *le traducteur passe son temps à aider l'ordinateur à comprendre le texte à traduire. L'étape de post-édition est de toute façon nécessaire...*

TransType: Le meilleur des deux mondes (Foster et al., 1997)

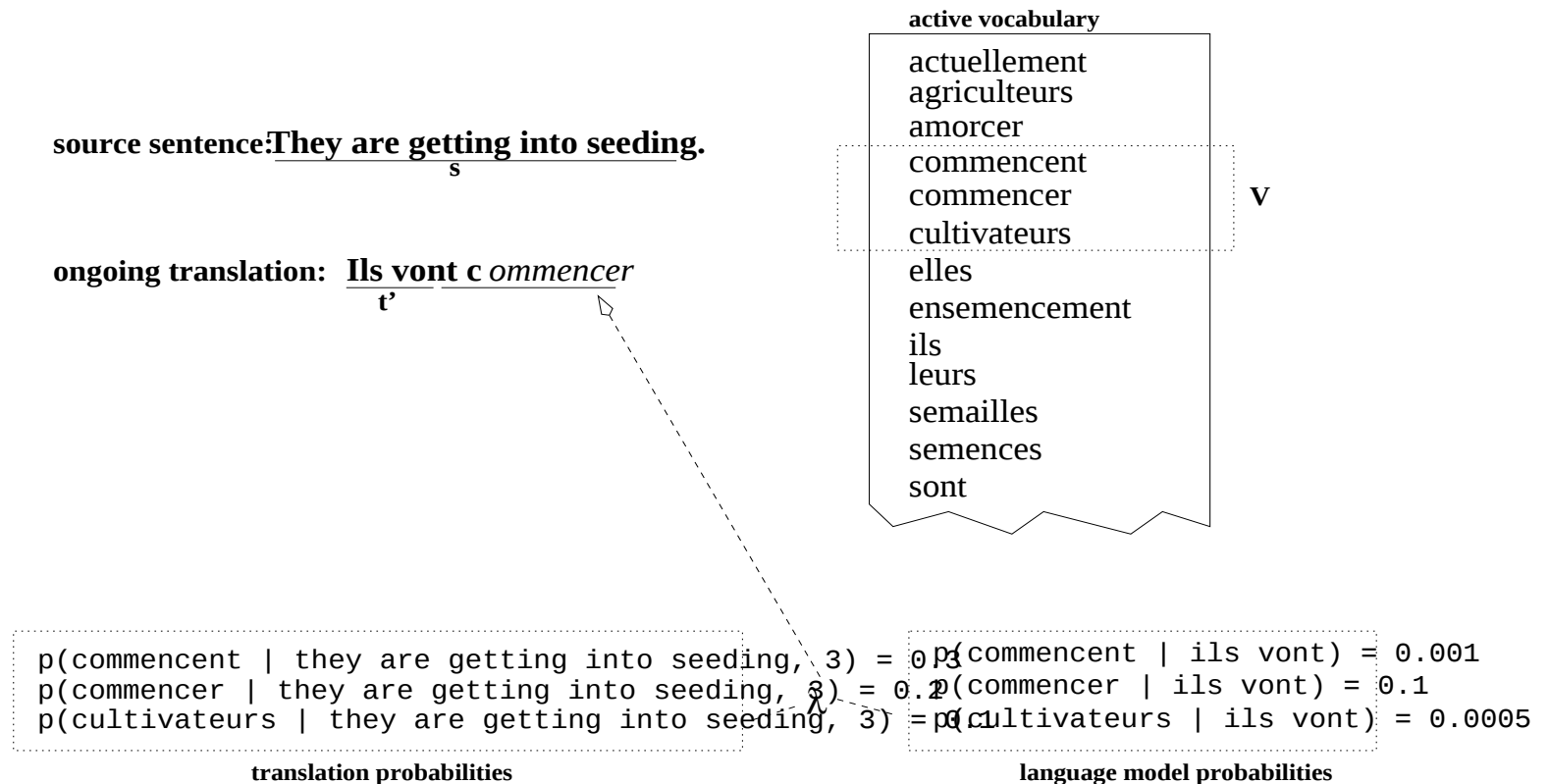
source sentence to be translated



proposed completions

ongoing target translation

Le cœur de TransType en un schéma



Well-established modeling techniques for both distributions:

Translation model	IBM2	(Brown et al., 1993)
Language model	interpolated trigram	(Jelinek, 1990)

Le cœur de TransType en équation

2 sources de prédiction: un modèle de langue et un modèle de traduction:

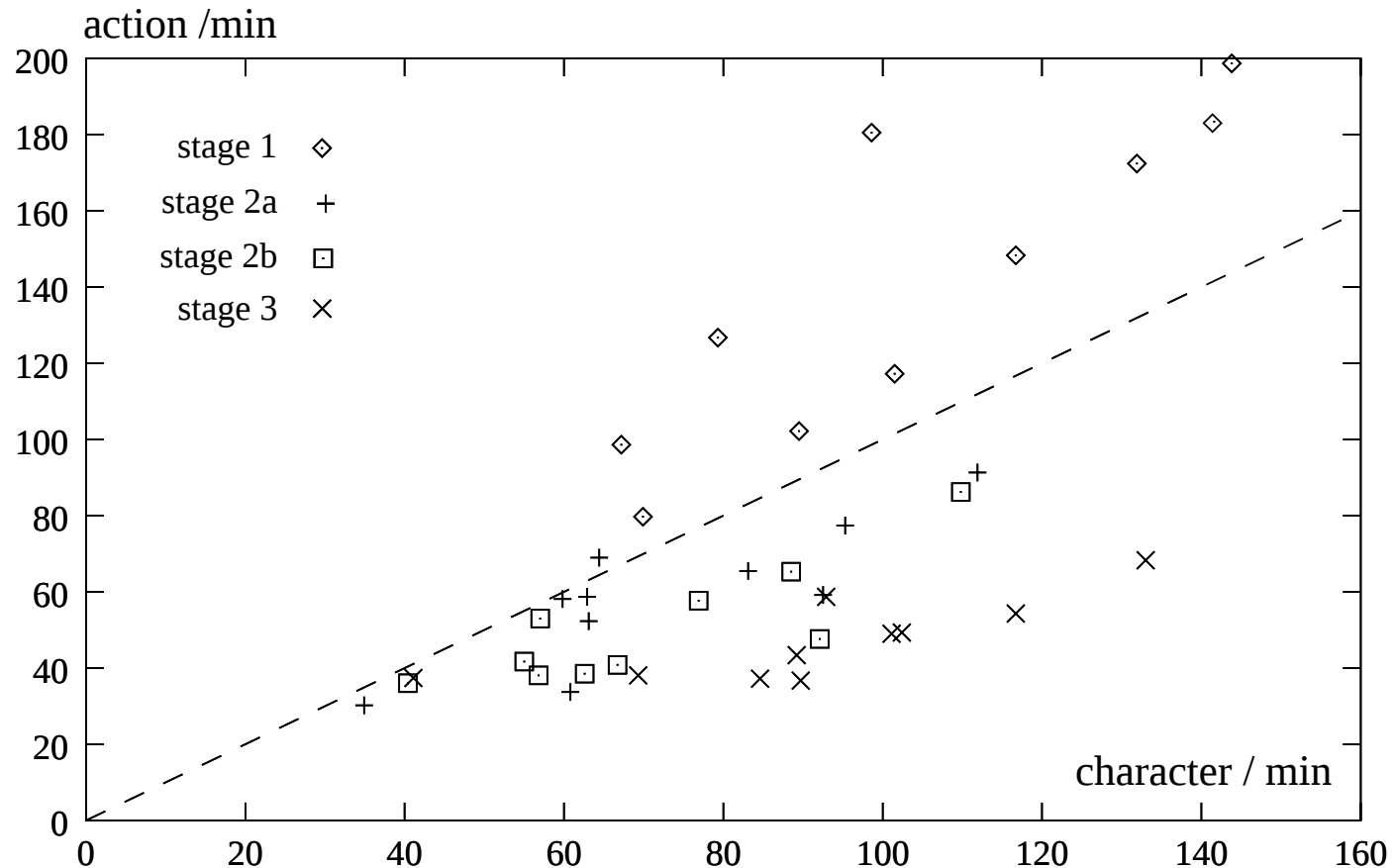
$$p(f_i | f_1^{i-1}, e) = \underbrace{p(f_i | f_1^{i-1}) \alpha(f_1^{i-1}, e)}_{\text{langage}} + \underbrace{p(f_i | e) [1 - \alpha(f_1^{i-1}, e)]}_{\text{traduction}}$$

$$\hat{f}_i = \operatorname{argmax}_{f_i \in V} p(f_i | f_1^{i-1}, e)$$

- Modèle de langue (cible): trigramme interpolé (Jelinek-Mercer)
- Modèle de traduction: IBM2
- λ : coefficients de pondération dépendants du contexte (f_1^{i-1}, e) (Langlais et al., 2000)

The spring 2000 *in situ* evaluation

Les utilisateurs de TRANSType semblent heureux, pensent que TRANSType les aide à traduire plus rapidement, mais qu'en est-il vraiment ?

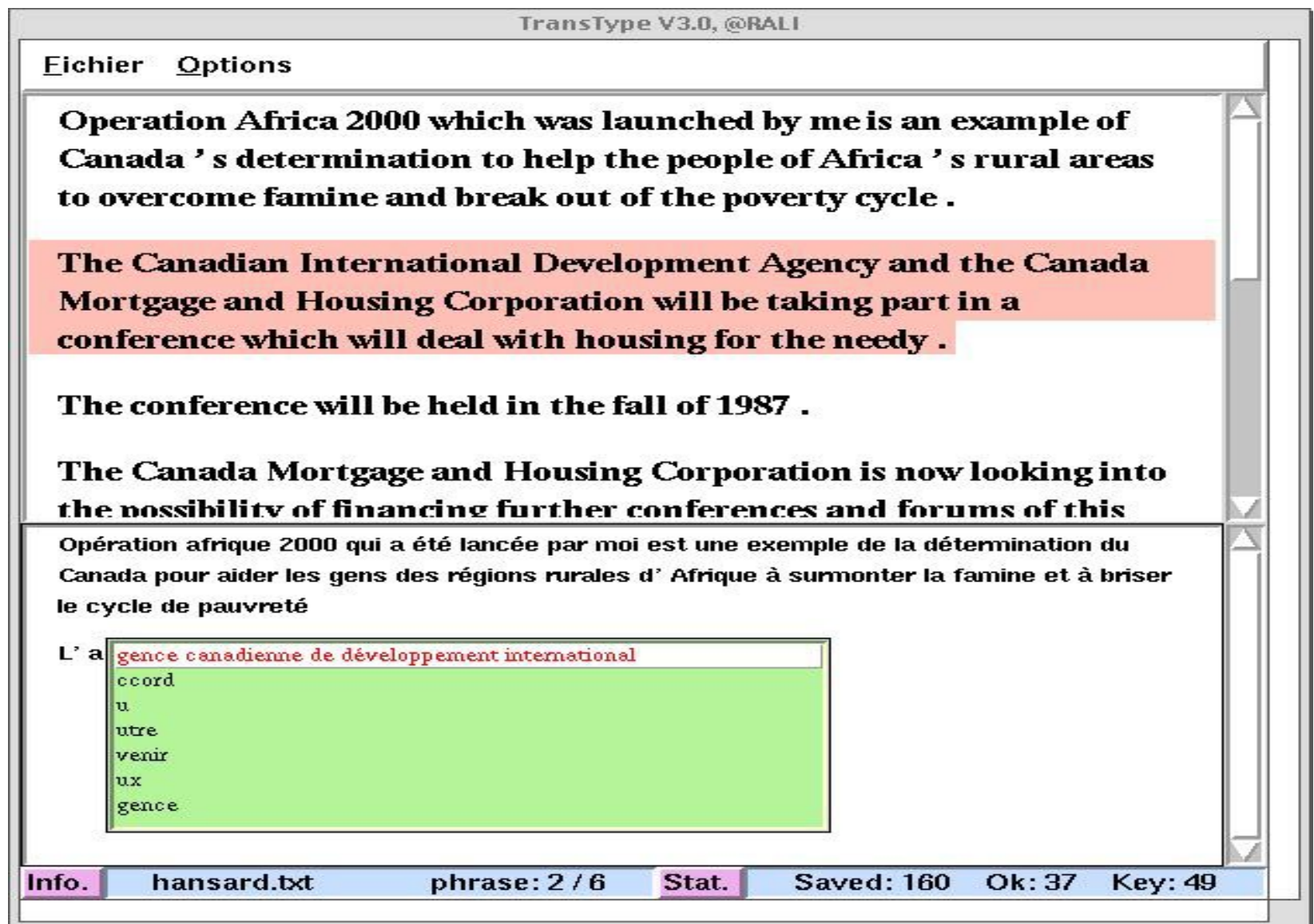


Why user lexicons would be of benefit

TRANSTYPE has been evaluated by ten translators who were asked to translate the same text (Langlais et al., 2000).

- Users like the tool **but** would really appreciate completions beyond the word level. User lexicons are likely to contain many terms (multiwords)
→ *this could help.*
- Lexicons play an important role in the translation process. Many translation companies employ terminologists to develop specialized lexicons.
- Lexicons are often the only means for a user to influence the translation engine.
→ *toward a lazy adaptative version of TRANSTYPE.*

TransType and user lexicons



Some problems with user lexicons

existential what is a user lexicon ?

→ *a list of source/target entries involving at least 3 words:*

<i>senior officials</i>	<i>hauts fonctionnaires</i>	<i>(4 words)</i>
<i>shipbuilding</i>	<i>construction navale</i>	<i>(3 words)</i>
<i>spokesman</i>	<i>porte-parole</i>	<i>(2 words) ← no !</i>

technical How should we integrate non probabilistic lexicons into a probabilistic engine ?

→ *a probabilistically minded hack could help*

ergonomic When and how should TRANSTYPE suggest target lexicon-entries ?

→ *right at the time the user needs it, in the popup menu*

availability Can we automatically extract a “user lexicon” from a given bitext ?

→ *not that obvious ! → first, we need to solve the existential problem*

Extrait d'un lexique bilingue acquis automatiquement^a

canadian association of university teachers	association canadienne des professeurs d'université
export enhancement pro- gram	programme de stimulation des exportations programme d'amélioration des exportations
farm debt review boards	bureaux d'examen de l'endet- tement agricole
federal cuts	compressions budgétaires fédérales
gender equality	égalité des sexes
love for canada	amour pour le canada
senior officials	hauts fonctionnaires
shipbuilding	construction navale

^aLes techniques utilisées seront vues dans *colloc.ps*

Log trace of a small session

Src.sent.:	<i>I shall return to this point in a few moments</i>	
Tgt sent.:	Je reviendrai sur ce point dans quelques moments	
In lexicon:	<i>few moments</i> → quelques minutes / quelques instants / quelques moments	
target tokens	typed	best completions in turn
Je	+	/Je
reviendrai	revi+	/le · r/etour · re/venir · rev/iens · revi/endrai
sur	+	/sur
ce	+	/ce
point	+	/point
dans	d+	/de · d/ans
quelques	que+	/le · q/uelques instants · qu/elques minutes · que/lques moments
moments	-	

$$saved = 100 - \frac{(8+7 \text{ validations}) \times 100}{41+7 \text{ spaces}} = 68.75\%$$

The spring 2001 *in situ* evaluation

We have carried out this summer a new evaluation of TRANSType in its user lexicon version. 9 translators were involved. Here are some results of the qualitative survey:

- What do you think about the possibility of including your lexicon into TRANSType ?
→ 9 answered: **Great !**
One added: would it be possible to add entries on the fly ?
- Would you be inclined to use TRANSType in your daily work ?
→ 6 answered: **yes**; 2 answered **maybe**.
Among the six happy users, 4 said it was because of the new functionality.

↪ (9 out of 10 translators answered yes during the 2000 evaluation)

Conclusions rapides sur TransType

- Une idée originale (Foster et al., 1997)
- Une étude de faisabilité concluante...
- L'idée séduit les utilisateurs...
- Un projet Européen (2002-2005) impliquant des partenaires industriels...
- Des clones de TRANSType commencent à voir le jour dans les laboratoires et dans des compagnies...
- Dictée sa traduction serait un plus (idée initiale du projet TransTalk)...
- Intégrer un véritable décodeur serait un plus.

Un concordancier bilingue interactif: TransSearch

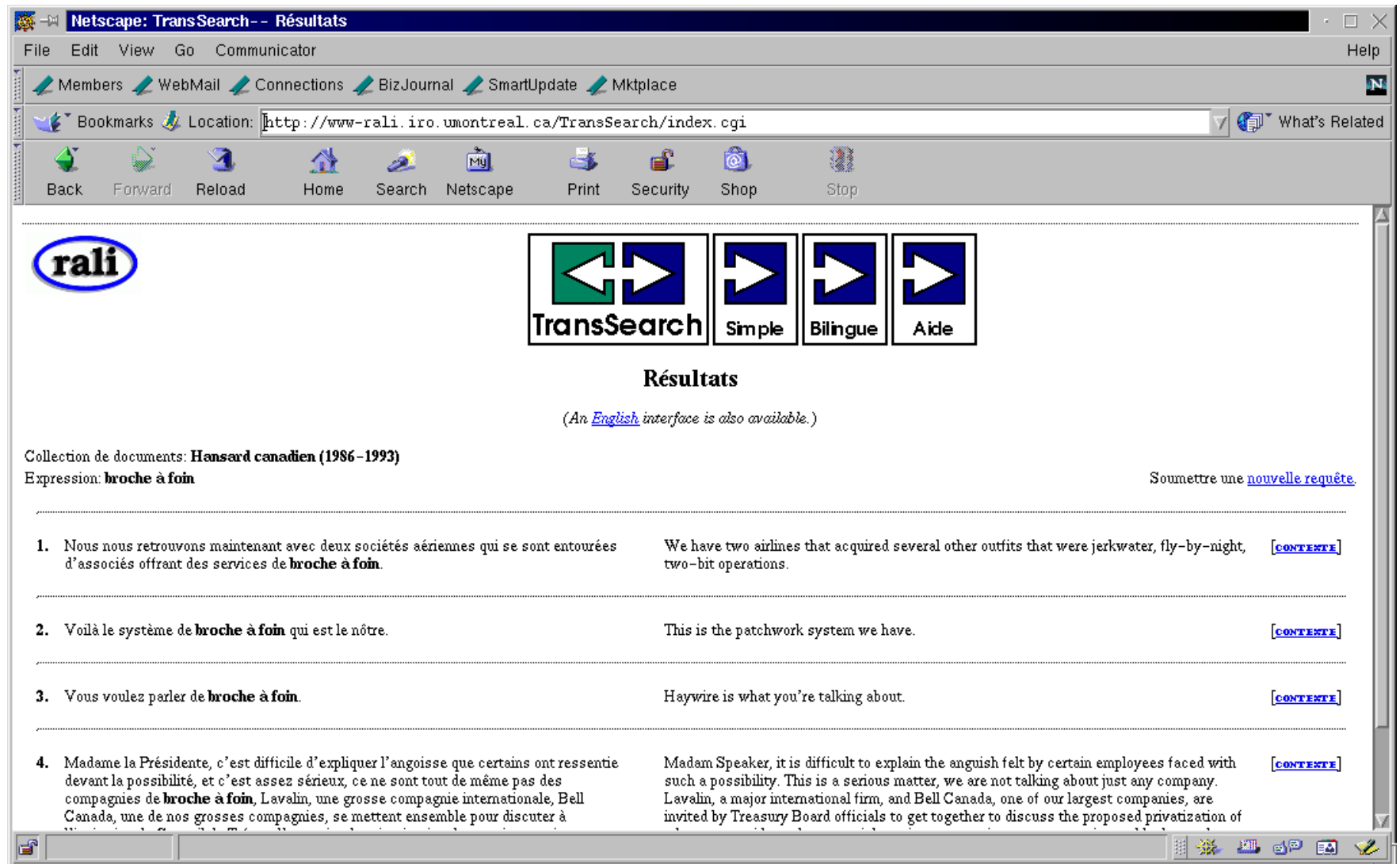
Citation du jour (Isabelle et al., 1993):

Les traductions existantes contiennent plus de solutions à plus de problèmes de traduction que n'importe quelle autre ressource.

- Un **concordancier** est un programme qui recherche des mots ou des expressions dans un corpus de texte. Lorsqu'on soumet une requête, le système fouille sa base de données (*mémoire*) et affiche toutes les occurrences trouvées, dans leur contexte.
↪ *un gros grep doté d'une interface conviviale*^a
- Un **concordancier bilingue** offre en plus l'avantage de retrouver et d'afficher la traduction de chaque passage qui satisfait la requête. De telles concordances bilingues sont rendues possibles grâce à des méthodes d'alignement qui établissent automatiquement des liens entre des segments correspondants de textes traduits.

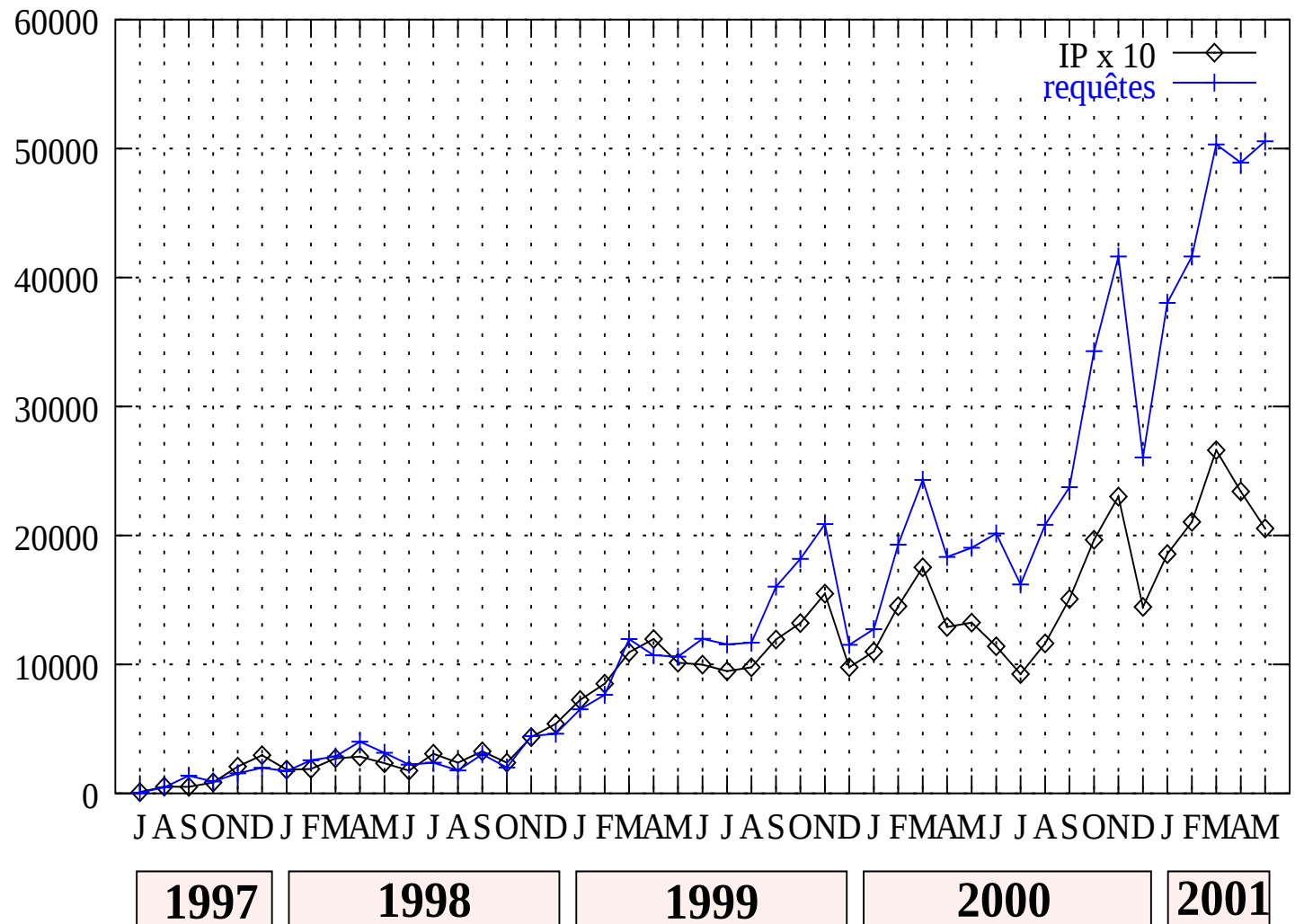
^aVoir par exemple <http://retour.iro.umontreal.ca/cgi-bin/lexiquum>

Un concordancier bilingue interactif: TransSearch



Un concordancier interactif: TransSearch

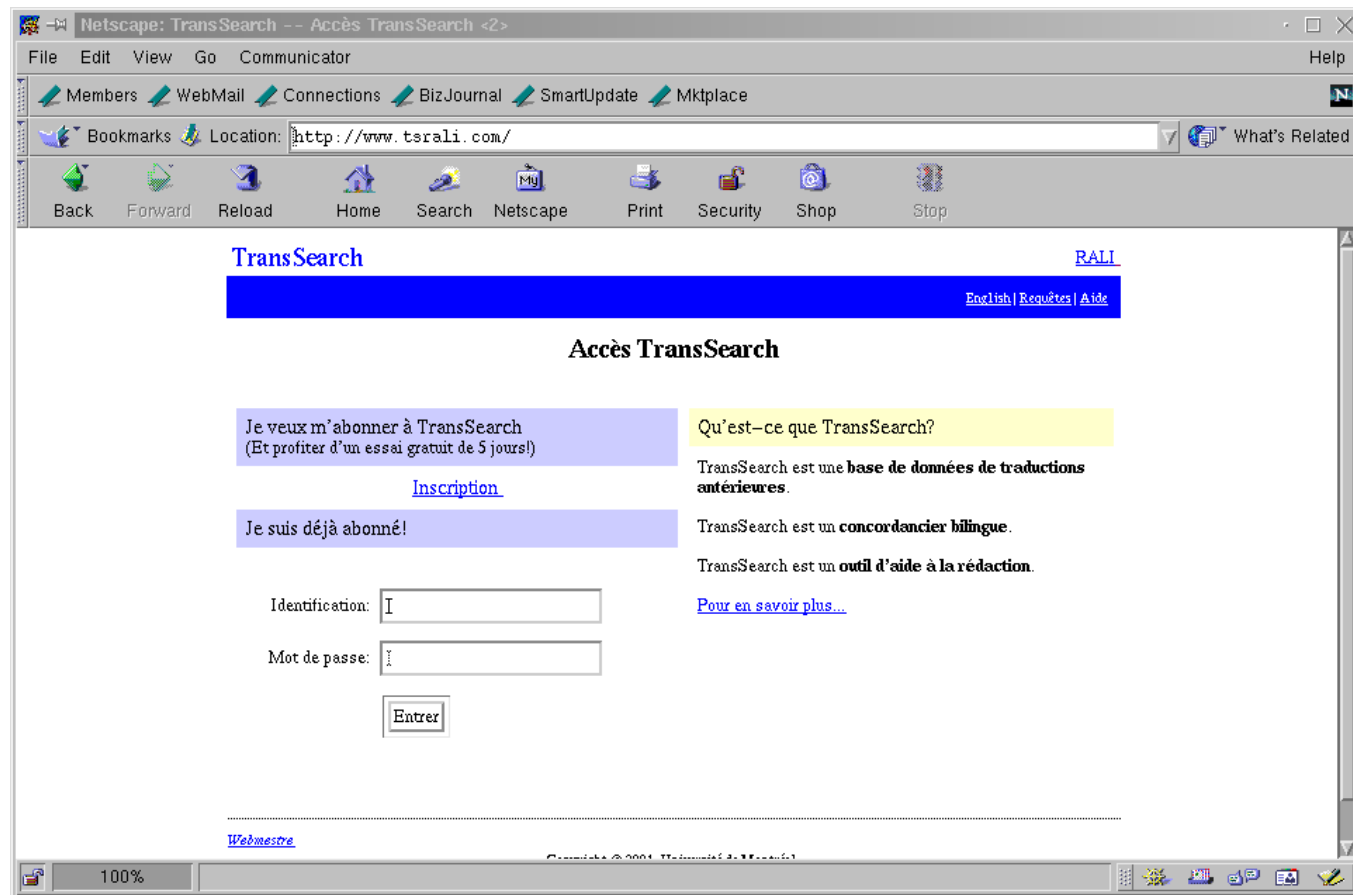
Statistiques d'accès



Un concordancier interactif payant (...)

TSRali.com

Fort du succès d'estime de ce service quotidiennement utilisé bien que jamais publicisé, le RALI décide d'arrêter le service TRANSSEARCH pour lui subsituer un service payant: TSRALI.COM



Un concordancier bilingue interactif: TSRali.com

TSRALI.COM offre à ses utilisateurs la possibilité de rechercher des solutions à leur problème de traduction dans deux bases de données (anglais/français):

Le Hansard: il s'agit du Journal des débats à la Chambre des communes du Parlement canadien. Notre corpus contient la totalité du Hansard, pour la période allant de avril 1986 à mai 2001. Il s'agit de traductions de l'anglais au français ou vice versa. **Le tout représente plusieurs dizaines de millions de mots.**

Les Cours canadiennes: ce corpus est constitué de documents provenant des collections de décisions de la cour suprême du Canada, de la Cour fédérale du Canada et de la Cour canadienne de l'impôt. Notre sélection comprend des documents datant de 1986 à nos jours.

Un concordancier bilingue interactif

Les requêtes reconnues par TSRali.com

verbatim Ex: *passer un sapin*

Le Québec, par la voix du
Bloc québécois, ne se laissera
pas **passer un sapin**.

Through the voice of the
Bloc, Quebec will make sure
it does not get taken.

dictionnaire Ex: *passer+ un sapin*

On **passe un sapin**, comme
on dit au Québec, aux autres
provinces.

They are pulling a fast one on
the other provinces.

Un concordancier bilingue interactif

Les requêtes reconnues par TSRali.com

Ellipse Ex: *former ... latter*

Le premier était assujetti à la réglementation, et l'autre pas.

The **former** was regulated and the **latter** was unregulated.

Bilingue Ex: *e=commitment, f=attachement*

Pour prouver son **attachement** à la démocratie, le premier ministre n'a qu'à dire aux Britanno-Colombiens qu'il nommera le sénateur qu'ils éliront.

To prove his **commitment** to democracy, the Prime Minister needs only to tell British Columbians that he will appoint the senator they elect.

Les mémoires de traduction (MT)

Survol

- Plusieurs systèmes existent: Trados (*Translator's Workbench*), IBM (*Translation Manager/2*), Atril (*Déjà-Vu*) et Star AG (*Transit*), MultiCorpora R&D inc. (*MultiTrans*), RALI (*TSRali.com*), etc.
- Les utilisateurs aiment ça (du moins ceux de *TSRali.com* :-)

Mais:

- niveau de granularité = fuzzy-phrase
 ↪ *révision de document, traductions routinières, etc.*
- les systèmes sous-phrastiques existants font appel au bon sens du ou des traducteurs pour peupler la mémoire
 ↪ *en combien de temps un tel outil s'avère productif ?*
 ↪ *que veut dire productif ?*

DÉJÀ-VU provides an unmatched breakthrough in productivity and consistency, wheter you are an in-house translation department, a translation solution provider or a freelance translator.

Motivations pour les MT

(Langlais and Simard, 2002)

- Vérifier jusqu'à quel point des exemples extraits **automatiquement** d'une mémoire de traduction peuvent être utiles à un traducteur (**alternativement à la traduction probabiliste**)
 - $\hookrightarrow$ *analyser la traduction (MT) est peut-être plus facile que de la générer (SMT)*
 - $\hookrightarrow$ *il est possible qu'une MT pallie les lacunes des modèles de traduction actuels (Marcu, 2001)*
- Une MT sur un domaine particulier peut être un moyen simple **d'adapter en ligne** un engin de traduction probabiliste (Carl and Langlais, 2002)
- Les modèles probabilistes ont du mal à gérer les **formes figées** (ex: **battre le fer pendant qu'il est encore chaud**), autrement que par énumération ou les **formes idiomatiques** (**kick the bucket**)

Mémoire de traduction (MT)

Qu'appelons-nous une MT dans cette étude ?

↪

Un (énorme) *bitexte* découpé et aligné au niveau des phrases

Ceci requiert:

- un corpus parallèle
↪ *les débats du Hansard, publiés entre avril 1986 et décembre 2001 (plus de 100 million de mots par langue).*
- un segmenteur en “phrases”
`sed -e 's/^\$/./g' $1 | tr -s "." " ." | sed -e 's/No\. *\[0-9\]/No \1/g'`
- un aligneur de phrases
↪ *Voir (Langlais et al., 1998) pour une comparaison de différents algorithmes; **en bref:** les systèmes “états de l’art” ont des taux de précision de l’ordre de 96% lorsqu’ils alignent des textes vraiment parallèles.*
- un indexeur

Mémoire de traduction (MT)

Qu'est-ce qui est nouveau alors ?

↪ **Le système d'interrogation de la mémoire (SIM)**

Ceci comporte:

- la **sélection** dans le texte à traduire de fragments à **rechercher** dans la mémoire (→ les requêtes),
↪ *Limitation de cette étude: nous faisons des requêtes verbatim de sous-séquences du texte à traduire*
- la **détection** de leur(s) équivalents dans la partie cible de la mémoire,
↪ *nous alignons automatiquement des paires de phrases au niveau des mots (à l'aide d'un modèle probabiliste)*
- le **filtrage** et la **présentation** du résultat des recherches à l'utilisateur (→ les briques traductionnelles)

Pour tous les “détails” concernant ces étapes, lire (Simard, 2003).

Sketch de la configuration testée

Le Tchoung-King^a du texte à traduire

The government is putting a \$2.2 billion tax on Canada 's most vulnerable industry, the airline industry



$\left[\begin{array}{l} [_{NP} \text{ The government }] [_{VP} \text{ is putting }] [_{NP} \text{ a \$2.2 billion tax }] [_{PP} \text{ on }] \\ [_{NP} \text{ Canada }] [_{NP} \text{ 's most vulnerable industry }], [_{NP} \text{ the airline industry }] \end{array} \right]$

↪ *par une cascade de modèles markoviens à la (Osborne, 2000):*

$$\begin{array}{ccc} w_i & \xRightarrow{HMM-1} & POS_i \\ (w_i, POS_i) & \xRightarrow{HMM-2} & IOB_i \\ (POS_i, IOB_i, POS_{i+1}, IOB_{i+1}) & \xRightarrow{HMM-3} & I\hat{O}B_i \end{array}$$

^aCorrection proposée par ispell au mot *chunking*

Sketch de la configuration testée

Sélection des requêtes

$[_{NP}$ The government] $[_{VP}$ is putting] $[_{NP}$ a \$2.2 billion tax] $[_{PP}$ on]
 $[_{NP}$ Canada] $[_{NP}$'s most vulnerable industry], $[_{NP}$ the airline industry]



The government / The government is putting / is putting / a \$2.2 billion
tax / a \$2.2 billion tax on / a \$2.2 billion tax on Canada / Canada 's
most vulnerable industry / 's most vulnerable industry / , the airline
industry / the airline industry



The government is putting / a \$2.2 billion tax on / Canada 's most
vulnerable industry / , the airline industry

Sketch de la configuration testée

Interrogation de la mémoire

The government is putting / a \$2.2 billion tax on / Canada 's most vulnerable industry / , the airline industry



Src: **the government is putting** things into law retroactively .
Tgt: le gouvernement prend l' habitude de légiférer rétroactivement .

...

Src: **the government is putting** certain social measures in place, such as this one.
Tgt: On met donc en place certaines mesures sociales, dont celle-là.

⋮

Src: Yes , **the airline industry** is an important industry .
Tgt: Oui , l' industrie aérienne est un secteur important .

Sketch de la configuration testée

Identification des équivalents traductionnels des requêtes

<i>Source sequence: The government is putting</i>	
11.80	le gouvernement est en train
12.96	, le gouvernement accule à la
12.99	le gouvernement donne de
<i>Source sequence: a \$2.2 billion tax on</i>	
69.60	une taxe de 2,2 milliards de dollars à
69.87	une taxe de 2,2 milliards
<i>Source sequence: Canada 's most vulnerable industry</i>	
16.49	industrie la plus vulnérable au
19.88	la plus vulnérable au Canada
<i>Source sequence: , the airline industry</i>	
10.96	, l' industrie aérienne
11.14	, de l' industrie aérienne
11.36	, l' industrie du transport aérien



matériel à présenter (après filtrages éventuels) au traducteur ou au système de traduction

Comment obtient-on les briques traductionnelles ?

Plusieurs approches possibles (Véronis and Langlais, 2000; Simard, 2003), ici alignement binaire descendant à la (Wu, 1997) qui fait usage d'un modèle IBM1 pour décider des points de coupure binaires:

[There] [is] [**a collective benefit**] [to secure] [air travel] .
[un avantage collectif] [est] [indéniable] .



[There] [is] [**a collective benefit**] [to secure] [air travel]
[un avantage collectif] [est] [indéniable]



[**a collective benefit**] [to secure] [air travel]
[un avantage collectif]



a collective benefit $\equiv$ un avantage collectif

Ressources

Chunker de l'anglais entraîné sur le Penn Treebank (*Wall Street Journal* – voir CONLL 2000 shared task)

Chunker du français entraîné sur une portion du corpus *Corfrans* (*Le Monde* - (Abeillé et al., 2000))

Mémoire de traduction les débats du Hansard publiés entre avril 1986 et décembre 2001

Modèles de traduction et de langue entraînés sur un bitexte de 1 639 250 paires de phrases du Hansard

Matériel de test: 1646 phrases (longueur moyenne de 20 mots) de transcriptions récentes du Hansard (mars 2002)

Quelques faits

- Environ 60 000 requêtes (d'au moins 2 mots) faites avec succès à la mémoire, 28 000 seulement avec un alignement potentiel
 ↪ *longueur moyenne des requêtes 4.6 mots (la plus longue était de 30 mots).*
 ↪ *Plus de 3 millions de matches, environ 500 000 retenus.*
- pour 11% des phrases à traduire, aucune requête (de deux mots ou plus) n'a trouvé de réponse
- moins de 4% des phrases ont été trouvées *verbatim* dans la mémoire
- environ 80% du matériel source à traduire était couvert par les requêtes
- le matériel cible ramené couvrait 62% de la *traduction oracle*
 ↪ *seulement 39% si on sélectionne les requêtes sources*

Mesure de la réutilisabilité d'une mémoire

Proposition 0: Couvertures sources et cibles

SEQUENCE:0	2	The fact				
#	mtf	mas	mss	mts	sep	mss
MATCH:25		3.31	1.00	le fait	===	the fact
MATCH:7	6.49	1.00	fait	===	the fact	
MATCH:3	6.49	1.00	la réalité	===	the fact	
MATCH:3	63.78	1.00	du	===	The fact	
MATCH:2	61.19	1.00	que	===	The fact	
MATCH:1	5.45	1.00	fait ,	===	The fact	
MATCH:1	6.02	1.00	En fait ,	===	The fact	
MATCH:1	7.30	1.00	un fait	===	the fact	
MATCH:1	7.30	1.00	a fait	===	the fact	
...						

SEQUENCE:0	3	The fact is				
#	mtf	mas	mss	mts	sep	mss
MATCH:24		5.53	1.00	le fait	est	=== the fact
MATCH:7	11.36	1.00	En fait ,	===	The fact is	
MATCH:6	8.35	1.00	Le fait	===	The fact is	
MATCH:6	9.76	1.00	La réalité , c'	est	=== The fact	
MATCH:3	120.24	1.00	,	===	The fact is	
MATCH:2	10.49	1.00	fait ,	===	The fact is	
MATCH:1	6.00	1.00	Le fait est que	===	The fact is	
MATCH:1	8.74	1.00	la réalité est	===	the fact is	
MATCH:1	11.13	1.00	fait	===	The fact is	
MATCH:1	12.05	1.00	Il est vrai que	===	The fact is	
...						

Mesure de la réutilisabilité d'une mémoire

Couvertures

Trois bitextes (100 paires de phrases dans chaque bitexte) assez différents de nature pour mesurer les taux de précision et de rappel obtenus: **Hansard**, le roman “De la terre à la lune” de Jules Verne et **Citi**, un rapport scientifique.

Langue source: anglais

Langue cible: français

Mémoire: Hansard disponible dans TSRALI.COM

bitexte	select-10 + cover			select-10 + cover + chunk		
	source	cible	ratio	source	cible	ratio
Hansard	95.8%	58.2%	0.6	74.3%	53.7%	0.7
Verne	71.9%	28.4%	0.4	45.9%	22.6%	0.5
Citi	86.2%	41.5%	0.5	58.2%	35.7%	0.6

Mesure de la réutilisabilité d'une mémoire

Proposition 1: Précision & Rappel

Soit $\mathcal{M}$ le matériel cible extrait de la mémoire et $\mathcal{G}$ l'oracle à reproduire:

$$precision = \frac{|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{M}|}, \text{ rappel} = \frac{|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{G}|}$$

$\hookrightarrow$ Mais calculer ces taux requiert un scenario utilisateur...

scénario	précision	rappel	f-mesure
cut&paste 1	55.5%	40.5%	46.8%
cut&paste 2	33.2%	24.2%	27.9%
cut&paste 3	23.6%	17.2%	19.9%
cut&paste 4	15.8%	11.5%	13.3%
paste-only	11.8%	8.6%	10.0%

yes-f1-sno-low-punc-sing-blind

Précision & Rappel

Exemple

Source: We have seen what has happened to the CanJets , the Canada 3000s and Royal Airlines .

Oracle: Nous avons vu ce qui est arrivé à CanJet , à Canada 3000 et à Royal .

Briques traductionnelles:

We have seen what has happened to	→	Nous avons vu quel sort a été réservé
	→	Nous avons vu
the Canada 3000s	→	Canada 3000
and Royal Airlines	→	et Royal Airlines

paste-only: $\mathcal{M} \cap \mathcal{G} = \{\text{Nous avons vu, Canada 3000}\}$
 $\hookrightarrow P=5/16, R=5/17$

cut&paste-1: $\mathcal{M} \cap \mathcal{G} = \{\text{Nous, avons, vu, Canada, 3000, et, Royal}\}$
 $\hookrightarrow P=7/16, R=7/17$

Mesure de la réutilisabilité d'une mémoire

Proposition 2: traduction probabiliste

L	$nb.$	SMT + MT			SMT		
		BLEU	WER	SER	BLEU	WER	SER
5	199	0.29	52.6	83.4	0.18	58.1	91.9
10	397	0.25	55.8	89.4	0.19	59.4	95.5
15	646	0.24	59.6	93.0	0.18	62.4	97.2
20	890	0.25	60.5	94.7	0.17	63.6	97.9
25	1026	0.24	62.2	95.4	0.16	64.6	98.2
30	1126	0.23	63.0	95.8	0.16	65.1	98.4

BLEU: décrit dans (Papineni et al., 2002) et fournit par NIST

WER: taux d'erreur au niveau des mots (distance d'édition où toute opération d'édition est pénalisée de la même manière)

SER: pourcentage de phrases qui ne sont pas identiques à la traduction oracle.

SMT + MT: exemples

Ce qui marche ~ bien

SRC the prime minister does believe in a weak dollar position .
REF le premier ministre croit dans un dollar faible .

does believe in	croit en
croit vraiment dans	elle croit en
il a réponse à	il croit à
the prime minister	le premier ministre affirme haut
au premier ministre	du premier
de s' enorgueillir	de se féliciter
du premier ministre	le premier ministre croit
le premier ministre	...

SMT le premier ministre n' en crois pas un dollar faible .
SMT+MT le premier ministre croit vraiment dans une position
de faiblesse du dollar

SMT + MT: exemples

Ce qui marche moyennement

SRC this report is nothing but a hodgepodge of motherhood statements .

REF ce rapport n' est qu' un ramassis d' évidences .

this report is nothing but	ce rapport
a hodgepodge	un tel fatras
une brochette	une série de
un tas	un méli-mélo
un salmigondis	un ensemble hétéroclite
of motherhood statements	de déclarations inattaquables
de vœux pieux	

SMT c' est un rapport sans fatras de déclarations inattaquables .

SMT+MT ce rapport, c' est un salmigondis de vœux pieux .

↪ *Des briques traductionnelles correctes ne s'assemblent pas nécessairement*

SMT + MT: exemples

Ce qui marche moyennement

SRC this is hardly a record to boast about .
REF il n' y a pas de quoi être fier .

this is hardly a record	c' est loin d' être un record
to boast about	de l' emploi dont
de quoi se péter	de quoi se vanter
de s' enorgueillir	de se féliciter
de se péter les bretelles sur	à se péter
il se vante de réduire	ou encore de le vanter
pour se vanter	s' en vanter
se pavane avec	se targue de l' autre côté de la chambre
se vantent	se vanter de
vraiment pas	à ce sujet parce

SMT c' est un peu les faits . >> . vanter
SMT+MT c' est loin d'être un record pour se vanter

⇨ *Problème de friction entre les briques traductionnelles*

SMT + MT: exemples

Ce qui marche moyennement

SRC with respect to the second part of his question , there is no contradiction .

REF quant à la deuxième question , il n' existe aucune contradiction .

there is no contradiction

il n' y a aucune contradiction

il n' y a pas contradiction

il n' y a aucune contradiction- je dis bien aucune-

aucune contradiction

il n' y a

il n' y a pas de contradiction

with respect to the second part of his question

quant à la deuxième partie de sa question

en ce qui concerne la deuxième partie de sa question

SMT ce qui concerne la deuxième partie de sa question , il y a aucune contradiction .

SMT+MT quant à la deuxième partie de sa question , aucune contradiction .

↪ *Le choix du matériel cible n'est pas toujours le meilleur*

SMT + MT: exemples

Ce qui marche mal

SRC what is his government doing to get it there ?
REF que fait son gouvernement pour que le dollar atteigne
 cette valeur ?

doing to get le cas pour rendre	font pour s'attaquer pour que pour tenir
it there il ne le soit pas tout à fait disposée	cela pose n' y investit
what is his government faire le gouvernement que va	est ce que lui et le gouvernement vont qu' est ce que son gouvernement quelle solution

SMT c' est son gouvernement se fait il ?
SMT+MT **qu' est ce que son gouvernement** **pour que** cela pose ?

↪ *Toute requête n'est pas bonne à faire*

Conclusions sur les MT

- Une première version d'un **véritable** SIM sous-phrastique
- Les bases d'une méthodologie d'évaluation d'un SIM
- Si l'on croît les scores BLEU, les exemples extraits de la mémoire ont un impact positif sur les traductions produites par notre moteur de traduction.
↪ en pratique, on a des problèmes de friction aux frontières des briques traductionnelles
- Place à de nombreuses améliorations (Simard, 2003):
↪ filtrages ad-hoc pour éliminer des requêtes improductives,
↪ d'autres méthodes d'alignements sous-phrastiques,
↪ des méthodes alternatives pour sélectionner les meilleurs alignements (réseaux de neurones)
- On peut (doit) faire beaucoup mieux que des requêtes *verbatim*
↪ les techniques EBMt peuvent être utilisées (Carl and Way, 2003)

Documents utilisés pour la rédaction de ces transparents

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for french. In LREC2000 (LRE, 2000).

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Ralf Brown. 1996. Example-based machine translation in the pangloss system. In *Proceedings of the International Conference on Computational Linguistics (COLING) 1996*, pages 169–174, Copenhagen, Denmark, August.

Michael Carl and Philippe Langlais. 2002. An intelligent terminology database as a pre-processor for statistical machine translation”. In *2nd International Workshop on Computational Terminology (COMPUTERM)*, pages 15–21, Taipei, Taiwan.

Michael Carl and Andy Way, editors. 2003. *Recent Advances in Example-based Machine Translation*. Kluwer Academic Publishers, Dordrecht, Netherlands.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text Mediated Interactive Machine Translation. *Machine Translation*, 12:175–194.

Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, François Perrault, Xiaobo Ren, and Michel Simard. 1993. Translation Analysis and Translation Automation. In *Proceedings of the 5th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Kyoto, Japan.

F. Jelinek. 1990. Self-organized language modeling for speech recognition. In A. Waibel and K. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann, San Mateo, California.

Philippe Langlais and Michel Simard. 2001. Récupération d’unités sous-phrastiques dans une mémoire de traduction. In *8e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 243–252, Tours, France, jul.

Philippe Langlais and Michel Simard. 2002. Merging example-based and statistical machine

- translation: An experiment. In *Proceedings of the fifth Conference of Association for Machine Translation in the Americas (AMTA)*, pages 104–114, Tiburon, California, oct.
- Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Montréal, Canada, August.
- Ph. Langlais, S. Sauvé, G. Foster, E. Macklovith, and G. Lapalme. 2000. Evaluation of transtype, a computer-aided translation typing system: A comparison of a theoretical- and a user- oriented evaluation procedures. In *LREC2000 (LRE, 2000)*, pages 641–648.
2000. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, June.
- Daniel Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 378–385, Toulouse, France, July.
- Miles Osborne. 2000. Shallow parsing as part-of-speech tagging. In *Proceedings of the 4th Computational Natural Language Learning Workshop (CoNLL)*, Lisbon, Portugal, September. ACL SigNLL.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA, July.
- Michel Simard and Philippe Langlais. 2001. Sub-sentential exploitation of translation memories. In *Machine Translation Summit VIII, Machine Translation in The Information Age*, pages 335–339, Santiago de Compostela, Galicia, Spain, September.
- Michel Simard. 2003. Mémoires de traduction sous-phrastiques. future thèse de l’UdeM, avril.
- J. Véronis and Ph. Langlais, 2000. *Evaluation of parallel text alignment systems: The ARCADE project*, volume 13, chapter 19, pages 369–388. Parallel Text Processing, Kluwer.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, September.